Remarks

This communication responds to the Office Action mailed August 18, 2008 for the application captioned above. The following remarks are respectfully submitted

§101 Rejection

Claims 16-18, 22-24, and 30 stand rejected under 35 U.S.C. 101 because the claimed invention is directed to non-statutory subject matter. In particular, the cited claims recite "software" without the enabling limitation of being provided on a computer readable medium. Appropriate correction in required.

§103 Rejection

Claims 1-5, and 10-18 stand rejected under 35 U.S.C. 103(a) as being unpatentable over Heckerman et al (6,260,011) and further in view of Nefian (7,165,029).

Claims 6 and 7 stand rejected under 35 U.S.C. 103(a) as being unpatentable over Heckerman et al and Nefian, and further in view of Okada et al (5,809,454).

Claim 8 stands rejected under 35 U.S.C. 103(a) as being unpatentable over the combination as applied to claim 6 above, and further in view of Wakamoto (6,283,760).

Claim 9 stands rejected under 35 U.S.C. 103(a) as being unpatentable over the combination as applied to claim 6 above, and further in view of Tanizawa et al (US PG Pub 2002/0030334).

Claims 19-21 stand rejected under 35 U.S.C. 103(a) as being unpatentable over Heckerman et al and Nefian, and further in view of Wakamoto.

Claims 22-24 stand rejected under 35 U.S.C. 103(a) as being unpatentable over Heckerman and Nefian, and further in view of Wakamoto.

Claim 25 stands rejected under 35 U.S.C. 103(a) as being unpatentable over Heckerman et al and Nefian, and further in view of Schultz (4,918,730).

Claims 26-28 stand rejected under 35 U.S.C. 103(a) as being unpatentable over Heckerman et al and further in view of Wakamoto.

Claim 29 stands rejected under 35 U.S.C. 103(a) as being unpatentable over the combination as applied to claim 26 above, and further in view of Nefian.

Claim 30 stands rejected under 35 U.S.C. 103(a) as being unpatentable over Wark (7,243,062)

Summary of our Application

In a simple example of applicants' invention, a user easily replaces the original voice of an actor or singer visible in an audio-video program, typically a short scene from a movie, with the user's own recorded and automatically lip-synchronized voice.

More generally, applicant's invention provides automatic replacing of a first audio stream, which contains sound events in sync with corresponding visual events in a moving picture signal, with a second audio stream that is automatically edited to match the timing of its sound events with those within the first audio stream and is played back automatically at the correct start time to thus be in synch with the video, thereby providing a convincing replacement for the first audio signal.

A preferred embodiment of applicants' invention has means and methods for

a) Specific, precise and necessary audio, video and other data preparation and storage,

b) Recording and then automated editing of the user(s) voice(s) based on comparing timing differences between detailed acoustic features of **two audio signals: for example, one being** the user's voice and the other being the original actor's or singer's voice (which is in lip-sync with the video images), and

c) the automatic insertion of the user(s) edited (lip-synced) voice(s) at the correct times during playback of the video stream, the playback of other selected audio streams, and the corresponding automatic removal of the original performer's voice during playback to enable the said automatic replacement of the original performer with the users voices.

Detailed review of the basis of the invention

In the original audio and video recordings, which are utilized by the invention, many sounds are already synchronized with corresponding visual events related to the generation of the sound. For example, speech or singing will be in synchronism with the visible mouth movements of a character (actor/singer), and sound effects, such as the slamming of a car door or beating of a drum, will be in synchronism with the corresponding visible actions in the moving picture. This has been state of the art in talking films since the 1930s.

What the present invention achieves is (in one instance) the replacement of a recorded original voiced utterance (synchronized with a corresponding moving image) with a new recording of a users' similar voiced utterance which, is processed and played back with the moving picture such that it: a) is in accurate synchronism with the original visible mouth movements of the relevant actor; b) is heard instead of the original voiced utterance, which is automatically suppressed; and c) any other audio heard with the original voice is heard in addition to the new voice. The invention can similarly achieve a synchronously timed replacement of a recorded original sound effect with a new recording of a similar but not identical sound.

In the prior art of Bloom and Marshall (US 4 591 928), a new audio signal recording (e.g. voice) having substantially the same timing pattern as an original audio signal recording could be made. This prior art result was achieved by editing a similar new audio signal to fit the timing pattern of features of the original audio signal. However, enabling the selection of sections and the replacement of the original recording with the corresponding new recording having timing such that the new recording would synchronize with visible movements in the original moving picture during playback had to be implemented manually by the user and required a number of steps. It will be further appreciated that two substantially similar patterns are not in register with each other until their respective corresponding boundary features, such as respective start and end features if the patterns are linear patterns, are made to coincide. Only then will all the corresponding features of the two patterns be in the same positions in space or time. If such

registration of the patterns is not achieved, there will be an offset between the two patterns, and where the patterns are sounds, they will not be synchronous.

After the prior art system of Bloom and Marshall had been used to edit the timing pattern of the new recording to match the pattern of the original recorded audio, a digital audio and video editing system, for example, could be used to allow a person to visually or aurally align the beginnings of the sound waveforms of the original and new recordings so that their respective corresponding internal audio features would occur, if played back, at the same instants in time, like two singers singing in unison. One of the two waveforms would have been shifted forwards or backwards in time to achieve this aligning. Then during a playback pass with the moving picture played in sync with the original audio, the editor/user would need to manually control suppression or muting of the original audio to allow the new recording to be heard as replacement audio. When these manual processes had been carried out the end result would be a precisely synchronized replacement of one audio portion as in the present invention, so that during a playback pass the new recording would produce sound that synchronized with the visible movements. These processes would typically be employed to replace dialogue recorded on location by dialogue recorded in the studio to create a final sound track.

The basic underlying principle is that if the new sound is synchronized to the old sound, then the new sound will be experienced by an observer as in synchronism with the visible action simply because the old sound was in synchronism with the corresponding action in the moving picture.

In a system according to applicants' present invention, the new replacement voice signal is edited automatically to have the same timing pattern of features as an original voice signal, and positioned automatically to play back with the corresponding moving picture image at the right time in accurate synchronism with the start, stop and internal audio (e.g. spectral energy) features of the original voice signal, and means are provided by which the original voice signal is automatically made inaudible at the right times, with the effect that the new voice signal seamlessly, automatically and synchronously replaces a selected portion of the original sound

track being played back with the moving picture. The selected portion will not usually include all the sounds occurring throughout the duration of the replacement, since the original concurrent background music and effects may be retained while only the voiced utterance is replaced. The present invention allows for the concurrent sounds not being replaced to be presented during playing of the section, making the illusion of true replacement complete.
In applicant's invention, the fine time-varying detail of one audio signal is electronically and then physically synchronized during playback of sound and picture. Sounds will be edited and output such that the user hears them at precisely the same time as he sees corresponding sound-creating visual events in the moving image. i.e. audio (that was created separately from the sound-creating visual events) is automatically synchronized to the moving image in a way that is observable by humans and that audio replaces original audio that was previously played synchronized with the moving image. It should be noted that both the original and replacement audio signals have their own, independent time references, and that synchronization means not only the detailed alignment of audio edited to match another signal's timing before playback with video, but also the actual synchronization achieved by playback of the edited audio started at the correct instant with a video stream that remains synchronized in the lip-synchronized sense during playback.

Applicants teach the automatic synchronization of one audio stream (e.g. voice) with another audio stream (e.g. voice) and the automatic playback of the replacement audio with a video stream at a precise start time (e.g. to achieve lip-synch). The presence of two audio streams is essential. The "replacing said selected portion" of one audio stream with another edited audio stream is also essential.

Summary of Heckerman et al

Heckerman et al (US 6,260,011, referred to hereinafter as "Heckerman"), provides "automated methods and apparatus for synchronizing audio and text data", where the audio specifically is a reading of the same words as the text. Heckerman gives a typical application for this: (Col. 1, lines 50-53) "For example, an individual attempting to improve his/her reading skills may wish to listen to the audio version of a book while having text corresponding to the audio being presented highlighted on a display device."

Heckerman clearly does not do any of the following:

- Provide a second new audio stream which is substantially similar to the first audio stream (which is lip-synchronized to a moving picture)
- Analyze the first and second audio streams and edit the second audio stream based on timing differences to achieve lip-sync to a moving picture provided by a video stream,
- Provide a means to playback the edited second voice as a replacement voice in lip-synch with the moving picture.
- provide audio (voice) replacement and playback of other synchronized audio streams

Clearly since there is no moving picture provided by a video stream in Heckerman, there is no audio lip-synchronized to a moving picture. **Furthermore, there are crucial differences between text data and a video stream that provides a moving picture.** For example, the lines of text in the box below can be read aloud in a variety of ways in time. In one extreme, a pause could be inserted between each spoken word. In another, the speech could be read continuously with no pause inserted between any words. The text could be read slowly or quickly. Thus, there is no unique time registration for text and any point in any sound or silence.

ALL THIS TEXT IS BEING DISPLAYED ON THIS PAGE AT THE SAME TIME. SO IS THIS TEXT.

On a display screen, all the words of the text in the box above can be displayed at the same time, or sequentially one word at a time, or several words at a time – but the words of the above text cannot be spoken all at the same time. There is no inherent timing in the text. This raises the following question: what is meant by a text display "synchronized" to the corresponding spoken words?

Heckerman is silent on how the "text data via the display" is presented "in a synchronized manner using the inserted pointer" (Claim 36). Heckerman does not describe the manner in which the text words are to be displayed while the audio files are played back.
Heckerman does not propose or teach that the displayed text has its own time reference before it is displayed.

The text data in Heckerman is given pointers that allow it to be displayed at a time during the playback of the corresponding audio words, but the time line is singularly that of the audio signal (Heckerman, Col 12, lines 13 – 31). Alternatively Heckerman creates a separate text file for each of a plurality of audio files (Heckerman, Col 13, lines 14 – 34). **Text, as discussed in Heckerman, cannot be an equivalent or a replacement for the second audio signal.** Firstly, the text data does not influence the time scale of the audio plus text display presentation. Secondly, since text data has no inherent timing features, there are no meaningful time differences between text data and corresponding audio and consequently no time difference function that is definable and measurable. Thirdly, there is no replacement of the audio signal in Heckerman by another audio signal. Fourthly, the text is not presented in sync with a moving picture video display.

Heckerman only deals with the "synchronizing" of text to a single audio stream. For example:

> "In the case of synchronized text and audio files, the computer system 120 can switch between audio and text presentation modes or simultaneously provide audio corresponding to text being displayed." (Heckerman Col 7, lines 52-56)

Heckerman further teaches that his method of "synchronization" is done by using speech recognition and silence position detection on the audio stream and then generating "pointers" "aligning" the text with the audio's silent intervals as described generally in the following claims:

> Claim 31. "A device for processing electronic text data and electronic audio data, comprising: a speech recognizer for performing a speech recognition operation on the audio data to produce a set of recognized text; means for globally aligning the recognized text with words included in the text data; means for identifying a first location in the recognized text where silence was recognized and where at least one correctly recognized word adjoins the recognized silence; and means for inserting into the text data, at the location aligned with said first identified location, a pointer to the audio data corresponding to the recognized silence." (Heckerman, Col 17 lines 5-19)
> Claim 36: "The device of claim 31, wherein the audio data and text data correspond to the same literary work, the device further comprising: a display; an audio output system; and means for simultaneously presenting the audio data via the audio output system and the text data via the display in a synchronized manner using the inserted pointer." (Heckerman, Col 18, lines 14-21)

What Heckerman means by "synchronizing" audio and text, can be seen from the following statement:

> "Since the pointer [110] is located from a content perspective at the same position in the audio and text files 804, 814, the pointer serves to synchronize the audio and text files in accordance with the present invention." (Heckerman, Col 12, lines 61-64.)

The phrase "located from a content perspective" highlights that the term and concepts of "location" and "same position" must take on different meanings in Heckerman depending on the type of content:

- In regard to audio, position is based on pointers which are data indicating a timing offset (from some defined start time) of silence starting or finishing before or after a word or group of words in the audio stream.

- In regard to text, the positions being pointed to are simply the stored text data corresponding to silence intervals between words in the audio stream.

Heckerman also states:

> (Col. 12, lines 41-51) "Various formats for the generated set 412 of synchronized audio and text files will now be described with reference to FIGS. 8-13. FIG. 8 illustrates a set 412 of synchronized text and audio files corresponding to the same exemplary literary work. In the FIG. 8 embodiment, a single text file 804, including time stamps, is used to represent the text corpus 802. A single audio file 814 which includes the same time stamps as text file 804, is used to represent the audio corpus 812. The dashed line between the text and audio files 804, 814 represents the linking of the audio and text files through the use of common time stamps."

Thus Heckerman indicates that his use of "synchronized" in the above context generally means corresponding audio and text data are simply "linked". In Heckerman, text is linked at time points to an audio stream by establishing a list of time stamps or "pointers" which somehow (since details are not disclosed in Heckerman) trigger the actual action of revealing the text word (or block of words) on a display screen while the corresponding audio is being heard. Heckermans' idea of synchronization of text to spoken words is not related to the detailed timing of spoken words but instead relies on the intermittent occurrence of silent intervals:

> "Since silence normally occurs at the ends of sentences and paragraphs, and it is these points which are of particular interest for the insertion of pointers for the purposes of text and audio synchronization, identifying points where silence occurs as potential locations where a pointer may be inserted is particularly useful." (Col. 11 lines 33 to 38)

In referring to Fig.7, Heckerman states:

> "Since the words preceding and following the silence were correctly recognized, the silence defined by audio time stamps 110, [time stamp] 117 would be identified in step 608 as **a point where the audio and text files should be synchronized**. **File synchronization** can be accomplished by, e.g., adding a bi-directional pointer linking the text and audio files, to the text files, and, optionally, to the audio files. By adding a pointer at the identified aligned location corresponding to silence, in this case the audio and text files will be synchronized at the start of a sentence." (Col. 11, lines 57)

Even if every recorded spoken word was separated from its neighbor by silence, which is not the case in natural speech, the maximum frequency of occurrences of text being "synchronized" (i.e. appearing on a visual display while the corresponding audio was playing) by Heckerman's method would be once a word. Heckerman requires that "the methods and apparatus be capable of synchronizing the **starting points of sentences and/or paragraphs in audio and text files with a high degree of accuracy**". This is all that is necessary with text, but would be far too coarse for analyzing an audio stream to provide lip-syncing. For lip-syncing, a short-term analysis is required which can extract the detailed acoustical features of speech and audio, and the feature measurements are generally made at a rate of around 100 Hz (see Bloom and Marshall).

Text in Heckerman cannot be used as an equivalent to or substitute for the second audio stream required in applicant's invention. Although **text** is sequentially ordered, it has no actual timeline of its own. Therefore text can have no unique timing relationship to any related audio. The aim of Heckerman is to produce text displayed "synchronized" to audio data - which in Heckerman means displaying a word or words in text form while the sound of the word or words is reproduced.

In Heckerman, only one audio stream is feature analyzed (to provide word recognition and positions of silence): there is no second audio stream to be analyzed.

The concept of "alignment" in Heckerman is shown in Fig.7 and described in Col.11, lines 3-38 as being achieved by determining start and end times of a recognized word in the audio stream and entering these times into a table along with the corresponding word of text. There is no second audio stream, no determination of timing differences between two audio streams, no editing of one audio stream to align the corresponding time varying features with the other, and no playback of an edited audio stream.

Figure 8 in Heckerman displays the heading "Synchronized Text & Audio" and shows one data storage unit for "Text Corpus" and one data storage unit with "Audio Corpus" showing simple a link between "Text with Time Stamp" in each storage unit – indicating the extent of the meaning of "Synchronized Text and Audio". i.e. there is no playback of data involved in this meaning of "synchronized".

Heckerman does not teach synchronizing a 2nd audio stream to a first audio stream and thereby to a moving picture. The examiner has referenced how in Heckerman text in a file is "synchronized" (different meaning) to audio by pointers to audio or text data. But in Heckerman there is no mention of any means for replacing one audio stream with another or for playing back an audio and moving picture video data. Furthermore, Heckerman cannot provide means or method for determining timing differences between an audio stream and a text file, because a text file has no inherent timing reference.

On page4, para.2, of the Action, the Examiner cites Heckerman at Col. 7, lines 54-5: "File **synchronization** can be accomplished by, e.g., adding a bi-directional pointer linking the text and audio files, to the text files, and, optionally, to the audio files" and notes that Heckerman "does not specifically disclose the receipt of a video file synchronized with an audio file". However, the video file, which provides a video stream is an essential element in Applicants' invention because

a. A first audio signal is synchronized with this video file and

b. A second audio signal is processed in order to be made to synchronize to and play back with this video stream.

Summary of and comparison with Nefian et. al. (US 7,165,029 B2)

Nefian teaches a method for improving the accuracy of speech recognition by combining the results of analyzing a video stream containing the face of a person speaking and deriving identifiable mouth patterns with the results of analyzing the videoed speaker's voice audio

stream, which is necessarily made at the same time as the audio is spoken. Nefian utilizes one audio stream which is inherently synchronized with the video stream but there is no second audio signal being analyzed or edited or played back. In Nefian, these are the only streams being analyzed and this is done only for the purpose of speech recognition. There is no second audio stream being analyzed for timing differences with this first audio stream, there is no audio editing, there is no playback of any edited audio with the video in synchronization or otherwise.

Thus, there is no "base for manipulating the timing of the audio stream to meet the goals of the invention" (Examiner's report, Page 4, para. 3). Applicants' invention requires a second new audio stream which is similar to the first audio stream (not present in Nefian), and requires analysis and manipulation of this second audio stream to synchronize the second audio stream with the first audio stream and thereby with the video data so that the second audio stream can be played back in sync with a moving picture. The purpose of the analysis in Nefian is to generate a stream of identified words with no timing reference, whereas the purpose of analysis of Applicants' invention is to generate an audio stream synchronized with a video stream in which the audio is not the original speaker but is played back to start at precisely the right time to be initially and continually in sync with the video signal.

The Examiner states on page 4, para. 3, of the Action: "As taught by Nefian, the receipt of synchronized audio and video streams is well known, and would have been obvious to included in Heckerman et al to one of ordinary skill in the art at the time of the invention". However, if the skilled person included a moving picture video stream synchronized to the single audio stream in Heckerman, the result would still only be text roughly "synchronized" (but obviously not with lip-synchronized accuracy) to the original audio and video stream; but there would be no replacement of the original audio with a new audio stream synchronized to the moving picture video stream and replacing the original synchronized audio stream during a playback phase. There would be no time difference computed, no editing, and no replacement of audio during playback of audio and video.

Summary of and comparison with Okada et al (US 5,809,454)

Okada et al states (Col. 3, lines 59-67) "It is a primary objective of the present invention to provide an audio reproducing apparatus capable of reproducing voices that are natural and comfortable to hear, even in a variable speed playback mode. It is another objective ... to provide a video/audio reproducing apparatus, which is equipped with an audio reproducing apparatus capable of reducing the time lag between the generation of voices and the movement of moving pictures, and a video decoder."

In Okada, an audio reproducing apparatus is described which can alter the speed of voice playback of an audio or audio plus video playback device (e.g. a MPEG video/audio reproducer) in response to the measured speed of the incoming (MPEG) coded data stream bit rate. The apparatus can work with an audio only coded data stream – or an audio and video coded data stream. In the latter case, the delay introduced by the audio processing can be compensated for by the video decoder delaying or quickening "the self-operating timing in accordance with the difference between the currently computed time [of position in the video playback] and the previously computed time" (Col. 11, lines 39-42) to keep the audio and video "synchronized".

The meaning of "synchronization" here is explained further: "This signal synchronization reduces the time lag between voices reproduced from the loudspeaker 24 and the movement of moving pictures on the display 22, so that the lip asynchronization falls within the allowable audible range of human beings." (Col. 11, lines 45-49). In other words, the original audio and video moving image which were in sync by virtue of being locked together as parts of the same data stream (and not by virtue of content) are both adjusted to remain "in sync" on average when the data stream rate is increased or decreased by manipulating the audio stream and ensuring corresponding video frames occur with the right portion of manipulated audio.

It should be noted that neither the audio nor the video stream data are being analyzed with regard to the timing of their content for the purpose of maintain synchronization. Any timing differences being detected are in effect, mechanical, based on position or timing indices (e.g. counted indices) or data processing speeds.

The timing "index signal" described as "information associated with time" is created and used in conjunction with a "current time" by a detector to determine "a signal delay time in the voice speed converting unit ... and supply "a signal indicating the signal delay time to the video decoder. The video decoder controls a self-operation timing based on the signal indicating the signal delay time".

Thus, in response to a change to the playback speed of an audio/video reproducer, the audio playback is sped up or slowed down (but kept more "natural sounding" in pitch and intervals) by Okada and, if video is present, the audio is kept closely in sync with the video. How close is not fully made clear but this is not relevant. The control of the rate of time compression/expansion is effectively controlled by the selected playback speed – as determined by the amount of data stored in a "ring memory 32".

The audio data rate and/or processing method is modified in order to remain in sync with a corresponding video program – but the control of this will not necessarily achieve lip sync. No second audio signal is being analyzed, edited or time compressed/expanded. No substitution of one audio signal for another during playback. "User can be provided with natural reproduced voices according to the playback speed." (Col. 10. Lines 51-52)

In Okada there is no detailed feature extraction for the purpose of pattern matching, no establishing of relative timing of detailed feature data, no comparison of feature data, no second audio stream, no analysis of a second audio stream, no editing of an audio stream, and no replacement of an audio stream with an edited signal as taught in Applicants' invention.

Comparison of Applicants' invention with Heckerman, Nefian and Okada is outlined by the Table below.

	(1) Data to be synchronized	**(2) Data to be synchronized to (1)**	**Automatic (lip) synch-ronized replacement of (2) by (1) enabled?**	**Applicants' Comments**
Applicants' invention	1st Audio signal e.g. user's recorded voice	2nd audio signal e.g. original performer's voice which is (lip-) synchronized to video (moving picture)	YES	
Heckerman et. al.	TEXT from books or literary work Synchronized has different meaning	Voice recording	NO	Text is not equivalent to audio or video data streams. No signals are replaced.
Nefian et. al.	None	None	NO	A video and audio stream taken of a person talking, and thus already inherently synchronized, are both analyzed and the results combined to improve speech recognition accuracy. Although the audio and video streams can be independently asynchronously processed (to recognize speech) and the results of each process (recognized words) combined for events in each stream at the same time, there is no manipulation of the timing of the audio stream.

	(1) Data to be synchronized	(2) Data to be synchronized to (1)	Automatic (lip) synch-ronized replacement of (2) by (1) enabled?	Applicants' Comments
Okada et. al.	Edited Audio stream	Video stream to which original audio was already synchronized	NO	During playback of an audio and video data stream at different playback speeds, an average re-Synchronization is enabled between a time- modified form of the audio stream which was initially synchronized with the video stream. During playback and audio processing, positional timing differences between the output modified audio stream and the video stream are computed and used to keep the reproduction of the two streams in sync to an observer.

Regarding the Examiner's comments on Applicants' Claim 1

Table of comparisons

	Applicants' wording	Examiner's reference	Applicants' observation
C1a	An audio and video data processor comprising:	Heckerman: "An audio and text data processor comprising"	Text and moving picture video data are different.
C1b	a selector for selecting at least a portion of an audio data stream, said audio data stream being synchronized with a video data stream;	(Heckerman, Col. 5, lines 41-42): " A plurality of M audio files 22, 24 form an audio corpus 20"	The audio files 22, 24 in Heckerman are NOT supplied already synchronized to a video data stream.
C1c	an audio feature analyzer for abstracting from said selected portion of said audio data stream a stream of time-varying features	(Heckerman, Col 11, lines 26-30): "locations in the recognized text where silence preceded and/or followed by	The "features" in the audio stream that Heckerman analyses and identifies are "locations ... where silence" preceded or followed recognized whole words. These are

	Applicants' wording	Examiner's reference	Applicants' observation
	and for abstracting corresponding time-varying features from an input audio data stream	correctly recognized words are identified. As discussed above, in the case of audio versions of literary works and other works read allowed [sic] and recorded for commercial distribution purposes, silence is often a particularly easy to recognize"	intermittently occurring features and consequently cannot be used for analyzing detailed timing differences between audio signals for waveform editing purposes. Text and timestamps cannot be analyzed for audio features, nor edited by a waveform editor, nor played back. We have modified our claim to emphasize the level of detail required.
C1d	a timing analysis and waveform editing processor adapted to determine timing differences between said stream of time-varying features and said corresponding time-varying features and to utilize said timing differences to edit said input audio data stream	(Heckerman, Col 11, lines 6-8): "matching recognized words or sequences of recognized words in the recognized text to those found in the text corpus"	"Waveform editing" to achieve synchronization between two audio data streams is clearly not the same as Heckerman's "matching" of recognized words in an audio stream to text data found in a text corpus. The waveform editing involves modifying the samples and or portions of one audio data stream to fit the same time pattern as in the other audio data stream and is not mentioned in Heckerman. Waveform editing of a second audio signal is based on having a first audio signal which provides a timing reference– and the signals are analyzed in sufficient detail to provide timing difference data. No timing difference data is described or created in Heckerman. "Matching" in Heckerman involves conceptual linking of a recognized spoken word with the text representation of that word. At best, the match creates a sequence and a point at which to display the word as text when the audio for that word plays. MODIFICATION TO CLAIMS – We will emphasize there needs to be two audio streams and distinguish them by describing a "first" and

	Applicants' wording	Examiner's reference	Applicants' observation
			"second" audio stream. Attention is drawn to the fact that this input stream is a second audio data stream which is substantially in sync with a first audio data stream. In Heckerman, there are no second streams in sync with a first one. Locations of silence are not time varying features suitable for comparing with another stream of time varying features for the purpose of editing audio. Location of silence pointers and recognized text are not representative of the audio waveform and cannot be considered an equivalent to "time-varying features of the input audio data stream".
C1e	a playback control module adapted to control running of said synchronized audio data and video data streams with said edited input audio data stream replacing said selected portion	(Heckerman, Col 11, lines 62-64): "File synchronization can be accomplished by, e.g., adding a bi-directional pointer linking the text and audio files, to the text files, and, optionally, to the audio files"	The simple "adding of bi-directional pointers" to data files is not the same as the complex processes of playback ("running") of edited audio data stream which is started and synchronized to another audio data stream, and thereby to a video data stream, and causing the replacement of a portion of the original audio signal to occur. In Heckerman, there is no editing of audio, no synchronization of audio, and no replacement of audio with audio during playback.

Regarding the Examiners Comments on Claim 2

The statement "Heckerman et al disclose a data processing system for audio and video data" is not correct. As indicated by the Examiner, "Heckerman et al disclose an audio and text data processor" (Action page 3, para 2). Applicants' observations on Examiner's comments are set out in the Table below.

	Applicants' wording	Examiner's reference	Applicants' observation
C2a	A data processing system for audio and video data, comprising	Heckerman	Text data and video data are different.
C2b	digitized audio and video data for providing an audio data stream synchronized with a video data stream	(Heckerman, Col. 5, lines 41-42) A plurality of M audio files 22, 24 form an audio corpus 20"	Heckerman completely omits the "video data stream", whereas synchronized moving picture video is essential to the purpose and function of Applicants' invention.
C2c	timing data representative of a plurality of selected times in a running of said synchronized audio and video data streams	(Heckerman, Col 8, lines 28-33) "The alignment module 318 is also responsible for aligning the audio and text files based on the identified alignment points, e.g., by inserting into the text and/or audio files time stamps or other markers which can be used as pointers between the audio and text files"	Heckerman's time stamps are only linking static text and audio running times and thus there is no reference to selected times in a corresponding synchronized video data stream. The "alignment module" in Heckerman "aligns audio and text files". In contrast, Applicants' invention aligns audio to audio.
C2d	audio feature data for providing a data stream of time-varying features abstracted from at least a selected portion of said audio data stream;	(Heckerman, Col 11, lines 26-30) "locations in the recognized text where **silence** preceded and/or followed by correctly recognized words are identified. As discussed above, in the case of audio versions of literary works and other works read allowed [sic] and recorded for commercial distribution purposes, silence is often a particularly easy to recognize"	There is a difference between the degenerate "feature" of a location of a section of silence in an audio stream on the one hand and, on the other hand, a stream of "time-varying feature data" that is suitably detailed in time and states (dimensionality) such that it can be used in a pattern-recognition process which can establish timing differences between two such audio steams from two audio signals. Detailed editing data cannot be derived based on the position of silence only. Heckerman does not use any silence or word location data to edit audio or to determine playback of an audio stream. At best, Heckerman's analysis will only provide times of starts and stops of words or blocks of concatenated words.

	Applicants' wording	Examiner's reference	Applicants' observation
C2e	an audio feature analyzer for abstracting a corresponding stream of time-varying features from an input audio data stream;	(Heckerman, Col 11, lines 26-30) "locations in the recognized text where silence preceded and/or followed by correctly recognized words are identified. As discussed above, in the case of audio versions of literary works and other works read allowed [sic] and recorded for commercial distribution purposes, silence is often a particularly easy to recognize"	Applicants' invention provides a feature analyzer that abstracts far more detailed and frequent time varying features from an audio stream and, more particularly, abstracts the same time-varying features from a second audio data stream similar to a first stream. Such second similar stream is not present in Heckerman.
C2f	a timing analysis and waveform editing processor adapted to determine timing differences between said streams of time-varying features and to utilize said timing differences to edit said input audio data stream and produce edited input audio data	(Heckerman, Col 11, lines 6-8) "matching recognized words or sequences of recognized words in the recognized text to those found in the text corpus"	The examiner has indicated that he equates "matching recognized words ...in the recognized text to those [words] found in the text corpus" with "a timing analysis and waveform editing processor adapted to determine timing differences between said streams of time-varying features". However, the text data used in Heckerman does not have time varying features – and therefore no timing differences can be generated between text and audio. Consequently, without timing differences, and, without the timing differences, no editing of audio data can be or is carried out in Heckerman.
C2g	a playback control module adapted to control running said synchronized audio data and video data streams with said edited input audio data replacing said selected portion.	(Heckerman, Col 11, lines 62-64) "File synchronization can be accomplished by, e.g., adding a bi-directional pointer linking the text and audio files, to the text files, and, optionally, to the audio files").	In Heckerman, "adding of bi-directional pointers" links the data files (viz. text data and audio data) to achieve "file synchronization", but there is no editing of audio, no synchronization of an audio stream to anything, and no replacement of audio with audio during playback. In contrast, Applicants' invention includes the complex processes of automated

	Applicants' wording	Examiner's reference	Applicants' observation
			playback, namely the running of edited audio which is started and synchronized to another audio and video stream and the replacement of a portion of the original audio signal.

Regarding Examiner's comments on Claim 3

Examiner states (Action page 6) "Heckerman et al disclose a data processing system comprising cueing data representative of timing of said selected portion of said audio data stream (Col. 8, lines 28 – 33 "The alignment module 318 is also responsible for aligning the audio and text files based on the identified alignment points, e.g., by inserting into the text and/or audio files time stamps or other markers which can be used as pointers between the audio and text files." As pointed out hereinbefore, Heckerman lacks essential feature of "A data processing system according to Claim 2". Furthermore, it should be noted that "cueing" has a normal definition, which we provide below:

From http://www.thefreedictionary.com/Cueing

tr.v. **cued**, **cu·ing**, **cues**

1) To give a cue to; signal or prompt.

2) To insert into the sequence of a performance: cued the lights for the monologue scene.

3) To position (an audio or video recording) in readiness for playing: cue up a record on the turntable.

Cueing data in Applicants' claim 3 is data which will be used to provide a user with a visual or audible signal or prompt to start their creation of replacement audio (e.g. start speaking) and additionally provide further "cues" as to when certain sounds in the first audio signal that is being replaced should be occurring. In Heckerman, there is data in the form of "time stamps or other markers which can be used as pointers between the audio and text files" but Heckerman does not describe or disclose how such data is used for cueing nor if it could be used for cueing. These time stamps are not intended to be used as "cueing data". Heckerman does not teach that

the "alignment module" provides a mechanism for utilizing the pointers as cueing data. Heckerman's data may represent starts or ends of words, or of sentences or of paragraphs, and there is no apparent distinction between these, indicating ever further that data in Heckerman is not used for cueing.

Regarding Examiner's comments on Claim 4

Examiner (Action page 7): Heckerman et al disclose: "a data processing system comprising additional digitized audio data for providing a further audio data stream synchronized with said video data stream. (Col. 8, lines 16-18) "The language model generation module 314 is used for generating, from a text corpus, a language model used by the speech recognizer 312". However, there is no suggestion in Heckerman that the "language model" is audio data, and consequently there is no basis for the Examiner's association of "The language model generation module ... used for generating, from a text corpus, a language model used by the speech recognizer" with Applicants' "a data processing system ... comprising additional digitized audio data for providing a further audio data stream synchronized with said video data stream." It should be noted that Heckerman (Col. 9 lines 18-22) says: "The language model 408 is generated by the statistical language model generation module 314 in such a manner, e.g., as a finite state network, so as to allow it to be combined with the acoustic model 410 in a straight forward manner for speech recognition purposes." and "The language model 314, is used to estimate the probability P(W) associated with a hypothesized sequence of words (W)." (Col. 10 lines 1-3). None of these further descriptions of language module 314 or speech recognition module 312 discloses "a data processing system comprising additional digitized audio data or a further audio data stream synchronized with said video data stream."

Regarding Examiner's Comments in Point 5 (p.15) on Claim 6 and 7

The Examiner rejects claims 6 and 7 "as being unpatentable over Heckerman et al and Nefian, and further in view of Okada et al (5,809,454)". However, the cited references fail to disclose essential elements of Applicants' invention, as explained in the Table below.

	Applicants' wording	Examiner's reference	Applicants' observations
C6a	storing digitized audio and video data for providing an audio data stream synchronized with a video data stream;	This claim element has been omitted by Examiner	It is essential that a first audio stream be provided that is synchronized with a moving picture video steam and that the synchronized data be stored.
C6b	storing timing data representative of a plurality of selected times in a running of said synchronized audio and video data streams	(Heckerman, Col 8, lines 28-33) "The alignment module 318 is also responsible for aligning the audio and text files based on the identified alignment points, e.g., by inserting into the text and/or audio files time stamps or other markers which can be used as pointers between the audio and text files."	In Heckerman, text is not synchronously streamable video data. Furthermore, "inserting a pointer" into text or audio does not provide "a means for selecting "a portion of a stream of the streamable data". The pointer only identifies a position.
C6c	selecting at least a portion of said audio data stream;	(Heckerman, Col. 5, lines 41-42) "A plurality of M audio files 22, 24 form an audio corpus 20"	Firstly, Heckerman does not mention a selector mechanism for streams. Secondly, the audio files in Heckerman are NOT synchronized to a video file nor are any of the "M audio files" in Heckerman described as a replacement for another audio file.
C6d	abstracting from the selected portion of said audio data stream audio feature data for providing a data stream of time-varying features;	(Heckerman, Col 11, lines 26-30) "locations in the recognized text where silence preceded and/or followed by correctly recognized words are identified. As discussed above, in the case of audio versions of literary works and other works read allowed [sic] and recorded for commercial distribution purposes, silence is often a particularly easy to recognize"	The "features" in the audio stream that Heckerman analyses and identifies are "locations ... where silence" preceded or followed recognized whole words. These are degenerate features and not suitable for analyzing detailed timing differences between audio signals for waveform editing purposes. Text and time stamps cannot be analyzed for audio features, nor edited by a waveform editor, nor played back.

	Applicants' wording	Examiner's reference	Applicants' observations
C6e	storing the abstracted audio feature data;	(Col 9, lines 1-5) "The speech recognizer module 312, generates from the audio corpus 20 a set 406 of recognized text which includes time stamps indicating the location within the audio corpus of the audio segment which corresponds to a recognized word."	The data being stored is "feature data" that is not provided in Heckerman.
C6f	storing an audio feature analyzer for abstracting a corresponding stream of time-varying features from an input audio data stream;	(Heckerman, Col 9, lines 1-5) "The speech recognizer module 312, generates from the audio corpus 20 a set 406 of recognized text which includes time stamps indicating the location within the audio corpus of the audio segment which corresponds to a recognized word."	Heckerman does not teach storing a feature analyzer capable of abstracting the time-varying feature required by Applicants' invention.
C6g	storing a timing analysis and waveform editing processor adapted to determine timing differences between said data stream of time-varying features and corresponding features abstracted from an input audio data stream;	[EXAMINER HAS OMMITTED THIS PORTION]	This is a significant part of the claim and supplies an essential function.
C6h	and storing a playback control module for controlling running said synchronized audio data and video data streams with edited input audio data from said processor replacing said selected portion	(Heckerman, Col 11, lines 62-64) "File synchronization can be accomplished by, e.g., adding a bi-directional pointer linking the text and audio files, to the text files, and, optionally, to the audio files"	Heckerman does not teach storing of "a playback control module" capable of running edited audio which is started and synchronized to another audio and video stream and causing the replacement of a portion of the original audio signal to occur. In Heckerman, there is no editing of audio, no synchronization of an audio stream to anything, and no replacement of audio with audio during playback.

The examiner (on p. 16, last para) states "Heckerman discloses the synchronization of audio to a text file as analyzed and discussed above" and notes that "synchronization of audio and a display is done" (Col 7, lines 54-57) ..." and continues citing Heckerman "computer system can ... simultaneously provide audio corresponding to text being displayed". However, Heckerman is silent how this display of text with audio is carried out. He does not say the text is stationary, or whether a highlight shows up words or blocks of words when the corresponding audio plays. However, the display of text in Heckerman (even if it were moving on a screen – a feature which is not disclosed in Heckerman) cannot be considered to be equivalent to a displayed video stream containing moving pictures which show sound producing events to which original and replacement audio are synchronized. The text has no inherent timing. The text in Heckerman is not text-word for spoken-word synchronized to the audio stream.

Examiner's assertions (Action page 17, para 2) that Nefian's teaching provides "a base for manipulating the timing of the audio stream to meet the goals of the invention" and that "as taught by Nefian, the receipt of synchronized audio and video streams is well known, and would have been obvious to include in Heckerman et al to one of ordinary skill in the art at the time of the invention" are traversed. Nefian neither discloses nor suggests any means for analyzing the time-varying features of a first audio stream in away that would enable a second audio stream to replace the first audio stream.

Examiner states (Action page17, para 4) that "Heckerman discloses a processor adapted to determine timing between two related streams of data" (Col. 13, lines 40-42 "pointers, synchronizing portions of the audio and text data which have been found to correspond to each other")" and goes on to state that Heckerman does not "specifically disclose an editing processor to determining timing differences between the streams." The reason Heckerman does not disclose this is because it is not possible to define timing differences between audio and text data. Heckerman himself, although he uses the terms "synchronizing portions of the audio and text data" describes the relationship simply as "portions of the audio and text which have been found to correspond to each other". Timing of the text is not mentioned as it does not exist. It is not possible to determine timing differences between the audio and text streams in Heckerman.

It is submitted that Examiner has misinterpreted Okada at the last para on page 17 of the Action. Okada teaches editing an audio stream to maintain normal voice pitch when the data stream is transmitted at a rate different from the normal data rate. Consequently Okada does not teach any technique that is relevant to the editing of time-varying features of a second audio stream to the same pattern of features in a first audio stream. Okada's process operates with a single audio stream. Applicants' invention requires two audio data streams.

On Page 18 the Examiner states (p. 18, paras 1 and 2) Wakamoto teaches "the storage of multiple channels of sound synchronized with a video data stream" and then states that "it would have been obvious ... to modify Heckerman et al in order to provide for multiple audio streams synchronized with the video stream." This conclusion is traversed since Heckerman never provides a moving picture video stream to which multiple audio streams – or indeed even one audio stream – can be considered synchronized. Heckerman only provides a text display.

Regarding Examiner's Comments on claim 7

On page 18, the Examiner has left out of his restatement of claim 7 key portions of the actual claim as re-supplied below in brackets and emboldened:

Examiner states "Heckerman et al disclose a method **[according to claim 6, further]** comprising the step of: storing cueing data representative of timing of said selected portion of said audio data stream. (Col. 8, lines 28-33) ..."

We refer to our previous commentary on the Examiner's comments on Claim 3 on both the omission of dependency and the other similar errors made on the understanding of "cueing data".

Regarding Examiner comments on claim 8

The Examiner rejection of claims 8 "as being unpatentable over the combination as applied to claim 6 above and further in view of Wakamoto (6,283,760)" (Action page 18, para 4) is hereby traversed since as has been explained hereinbefore, the cited references do not disclose or suggest a combination that provides Applicants' invention. Furthermore, Wakamoto does not teach concurrently starting audio data.

Regarding Examiner's comments on Claim 9

Tanizawa does not overcome the deficiencies of the combination of prior art cited by the Examiner and is not an obvious addition. In Tanizawa, as shown in Fig. 79 and described in Paragraph 0420 cited by the Examiner, the gain control is applied to overlapping end and beginnings of waveform pitch periods during time expansion and compression of waveforms. This positions of these gain controls is dependent on details of two waveform portions – which is not at "selected times" as stated in our claim. Instead, Applicants provide predefined selected positions at which the user controls the option of whether the gain control is applied at the selected boundaries of larger blocks (e.g. sentences) of audio waveforms in order to enable the goal of audio replacement of one signal by another.

Regarding Examiner's comments on Claim 10

The Examiner has stated "Heckerman et al disclose a method of processing audio data, comprising the steps of:

	Applicants' wording	Examiner's reference	Comment on Examiner's comments
C10a	providing an original audio data stream synchronized with a video data stream	None – This claim portion has been omitted by examiner.	It is essential to note that the original audio stream is already synchronized with a video data stream.
C10b	selecting at least a portion of said original audio data stream	(Heckerman, Col. 5, lines 41-42) "A plurality of M audio files 22, 24 form an audio corpus 20"	Heckerman's audio files are NOT synchronized to a video file. Also, there is no audio in the corpus to provide a second, replacement audio stream.
C10c	storing an input audio data stream substantially in synchronization with a portion of said video data stream corresponding to the selected portion of said original audio data stream	(Heckerman, Col. 9, lines 1-5) "The speech recognizer module 312, generates from the audio corpus 20 a set 406 of recognized text which includes time stamps indicating the location within the audio corpus of the audio segment which corresponds to a recognized word."	Applicants' invention requires storage of an input (second) audio data stream with properties similar to a first audio stream that already synchronizes with a moving picture video. Heckerman's "recognized text" from a stored audio signal and time stamps is not an equivalent. Nor is Heckerman's audio synchronized to moving picture video. Text and timestamps cannot be analyzed for audio features, nor edited by a waveform editor, nor

	Applicants' wording	Examiner's reference	Comment on Examiner's comments
			played back.
C10d	abstracting from said input audio data stream a stream of time-varying features of the input audio data stream;	(Heckerman, Col 11, lines 26-30) "locations in the recognized text where silence preceded and/or followed by correctly recognized words are identified. As discussed above, in the case of audio versions of literary works and other works read allowed [sic] and recorded for commercial distribution purposes, silence is often a particularly easy to recognize"	It is essential to note that "said input stream" is a second audio data stream, which will be brought into sync with a first audio data stream. In Heckerman, there is no second stream. Locations of silence are not suitable time varying features for comparing with another stream of time varying features for the purpose of editing audio. Location of silence pointers and recognized text are not representative of the audio waveform and cannot be considered an equivalent to "time-varying features of the input audio data stream". See also **C1c** and **C2d.**
C10e	comparing the abstracted stream of time-varying features with a corresponding stream of time-varying features abstracted from said selected portion of said original audio data stream and determining timing differences between said streams of time-varying features;	(Heckerman, Col 11, lines 6-8) "matching recognized words or sequences of recognized words in the recognized text to those found in the text corpus"	In Heckerman, there are not two streams of time-varying features to compare because there are not two corresponding audio data streams. Moreover, there are no means or method of determining timing differences provided in Heckerman. Nor can time differences be generated between an audio stream and a text file which has no inherent timing. See also comments on "matching" in **C1d** and **C2f**
C10f	utilizing said timing differences to edit said input audio data stream and produce edited input audio data	(Heckerman, Col 11, lines 62-64) "File synchronization can be accomplished by, e.g., adding a bi-directional pointer linking the text and audio files, to the text files, and, optionally, to the audio files") ; and	The "adding of bi-directional pointers" to data files (one being text data and the other being audio data) to achieve "file synchronization" is significantly different from "utilizing … timing differences" between the signal being edited and another similar signal, to edit the [second] input data stream and produce edited audio data. In Heckerman there is no timing difference data, no second (input) audio data stream to edit based on timing differences, no waveform editing method or means, and no edited waveform.

	Applicants' wording	Examiner's reference	Comment on Examiner's comments
C10g	running said synchronized original audio data stream and video data stream with said edited input audio data replacing said selected portion	(Heckerman, Col 11, lines 62-64) "File synchronization can be accomplished by, e.g., adding a bi-directional pointer linking the text and audio files, to the text files optionally, to the audio files."	In Heckerman, there is no editing of audio, no synchronization of a second audio signal to a first original audio signal, and no replacement of a first original audio signal with a second audio signal during playback. See also comments in **C1e.**

Regarding Examiner's Comments on Claim 12

The Examiner states that "Heckerman et al disclose a method according to claim 10 wherein more than one portion of said original audio data stream is selected" citing Heckerman "pointers, synchronizing portions of the audio and text data" (Col. 13, lines 40-42). Applicants have commented above on the deficiencies of Heckerman in relation to Applicants' claim 10. Furthermore, the audio data stream being referenced in Applicants' claim 12 is a first, original audio stream which is already synchronized to a video stream.

Regarding Examiner's Comments on Claim 14

	Applicants' wording	Examiner's reference	Applicants' observation
C14a	means for deriving from audio data feature data representative of audible time-varying acoustic features of the audio data	(Heckerman, Col 11, lines 26-30) "locations in the recognized text where silence preceded and/or followed by correctly recognized words are identified. As discussed above, in the case of audio versions of literary works and other works read allowed [sic] and recorded for commercial distribution purposes, silence is often a particularly easy to recognize"	Neither the "recognized words" in Heckerman, nor the "locations ... where silence" are preceded or followed recognized whole words, are suitable for analyzing detailed timing differences between audio signals for waveform editing purposes. At best, Heckerman's analysis will only provide times of starts and stops of words or blocks of concatenated words.
C14b	means for selecting from data representing	(Heckerman, Col 11, lines 39-[4]2) "for a pointer to	In Heckerman, neither text nor any other stream can be considered

	Applicants' wording	Examiner's reference	Applicants' observation
	synchronously streamable video and audio data representing a portion of a stream of the streamable data and measuring durations of and intervals containing audible time-varying acoustic features of the audio data	be inserted into the text and/or audio for synchronization purposes, the recognized text, bracketing the identified point of silence must have been correctly identified "	"synchronously streamable video data". Furthermore, "inserting a pointer" does not fully provide "a means for selecting "a portion of a stream of the streamable data". The pointer only identifies a position.
C14c	means for populating a database with data and measurements provided by said selecting and measuring means	(Heckerman, Col 9, lines 1-5) "The speech recognizer module 312, generates from the audio corpus 20 a set 406 of recognized text which includes time stamps indicating the location within the audio corpus of the audio segment which corresponds to a recognized word."	The data being populated is "feature data" that is not provided in Heckerman, as observed above in C14a.

Regarding Examiner's Comments on Claim 15

The Examiner states, "Heckerman et al disclose an apparatus [...]comprising means for populating said database with text related to said data and measurements provided by said selecting and measuring means (Col. 8, lines 14-16 "the speech recognizer module 312 generates a set of recognized text with time stamps from one or more audio files"). However, Applicants' claim 15 is defining apparatus "according to claim 14" and consequently Heckerman fails to disclose or suggest such apparatus, as explained above in relation to claim 14.

Regarding Examiner's Comments on Claim 16

Examiner's rejection of claim 16 is traversed for the same reasons given above in relation to claims 1, 2, 6 and 10. In particular also regarding:

- "a feature analysis program … " refer to Applicants' observations at C10d, C1c and C2e.

- "a comparison and timing program ..." refer to Applicants' observations at C10e, C1d and C2f.
- "an editing program ... and" refer to Applicants' observations at C10f, C1d and C2f.
- "a streaming program..." refer to Applicants' observations at C10g, C1e and C2g.

Regarding Examiner's Comments on Claim 17

Examiner's rejection of claim 17 is traversed for the same reasons given above for claim 14. In particular: "a feature analysis program … " refer to our response in comments C14a.

- "a selection and measuring program ..." refer to Applicants' observations at C14b.
- "a database program … " refer to Applicants' observations at C14c.

Regarding Examiner's Comments on Claim 18

Examiner's rejection of claim 18 is traversed for the same reasons as his rejection of claim 17.

Comments regarding Examiner's Item 8 (p. 19) and claim 19-21

Examiner's rejection of claims 19-21 "as being unpatentable over the combination of Heckerman, Nefian and Wakamoto is traversed for the following reasons. Heckerman on its own and in combination with Nefian and Okada are all irrelevant as prior art as described previously above. The Examiner states (Action page 20 para 2) that "Heckerman et al disclose an apparatus for processing audio and video data" which is not the case since Heckerman relates to audio and text. Examiner continues saying Heckerman "notes that synchronization of audio and a display is done" - and quotes Heckerman as saying (Col. 7, lines 54-57) "In the case of synchronized text and audio files, the computer system … can simultaneously provide audio corresponding to text being displayed"- but does not describe how or in by what means this is achieved. Examiner (Action page21, para 1) then states how "Nefian teaches the receipt of a

video stream and a synchronized audio stream … providing a base for manipulating the timing of the audio stream to meet the goals of the invention". This is incorrect as explained previously in our "Summary of and comparison with Nefian" and regarding claim 6 . Similarly it is incorrect to state (Action page21 para 2) "As taught by Nefian, the receipt of synchronized audio and video streams is well known … and would have been obvious to include in Heckerman...".

Regarding Examiner's comments on Claim 20

In addition to Applicants' comments on the rejection of claim 19, it should also be noted that Heckerman requires a speech recognition system which, as he admits, can make errors: (Col 3, lines 46-55) "Unfortunately, with known speech recognition techniques, recognition errors occur. In addition, even when recognition errors do not occur, differences may exist between an audio and text version of the same work due, e.g., to reading errors on the part of the individual or individuals responsible for generating the audio version of the work.

The present invention uses a combination of silence detection and detection of actual words for purposes of synchronizing audio and text versions of the same work. " Heckerman does not provide "scene data". Thus, Heckerman's recognition technique does not populate "said database with text related to said scene data and measurements".

Regarding Examiner's comments on claim 21

The Examiner states (Action page 22, para 2) that "As taught by Wakamoto, the storage of still data is well known", and also that the "still data … provides the user with a means of accessing various locations in the recording, and would therefore have been an obvious addition to Heckerman". Applicants do not claim that still data "provides a means of accessing various locations in the recording".

Comments Regarding Examiner's Item 9 and claims 22 to 24

The Examiner states (Action page 22, para 4) "Regarding claim 22, Heckerman et al disclose audio and video data processing software..." However, Heckerman does not disclose "video data processing" and discloses instead "an audio and text data processor". Text is neither audio nor video and the synchronization of text to audio is not relevant. In his rejection of claims 22 to 24 the Examiner cites the same or similar Heckerman passages as he did in his remarks on claims 17, 10 and 14. Applicants refer to Applicants' previous comments correspondingly in claims 10, 14, 15, 17 and 18.

Comments Regarding Examiner's Item 11 and claims 26 to 28

A method of processing audio data, comprising the steps of:

	Applicants' wording	**Examiner's reference**	**Applicants' observation**
C26a	selecting from data representing synchronously streamable video and audio data scene data representing a portion of a stream of the streamable data	(Heckerman, Col. 11, lines 39-[4]2) "for a pointer to be inserted into the text and/or audio for synchronization purposes, the recognized text, bracketing the identified point of silence must have been correctly identified "	Heckerman does not contain data representing both synchronously streamable video and audio data; the cited passage does not provide a "selection". The examiner fails to take account of the fact that this input stream is a second audio data stream.
C26b	measuring durations of and intervals containing audible time-varying acoustic features of the audio data	(Heckerman, Col. 11, lines 26-30) "locations in the recognized text where silence preceded and/or followed by correctly recognized words are identified. ..."	The "features" in the audio stream that Heckerman analyses and identifies are degenerate features and not suitable for analyzing detailed timing differences between audio signals for waveform editing purposes.
C26c	zpopulating a database with scene data and measurements selected from and measured in the scene data.	(Heckerman, Col. 9, lines 1-5) "The speech recognizer module 312, generates from the audio corpus 20 a set 406 of recognized text which includes time stamps indicating the location	Heckerman determines where silence is between words or blocks of words without regard to whether these silences correspond to "scene data and measurements".

	Applicants' wording	Examiner's reference	Applicants' observation
		within the audio corpus of the audio segment which corresponds to a recognized word."	

With regard to the combination of additional passages the Examiner cites on Action page 29 that are taught from Nefian and Wakamoto, none of the cited references, individually or in combination provide the "method of processing audio data" of Applicants' invention, as explained hereinbefore.

Regarding Examiner's comments on claims 27, 28 and 29

For reasons given hereinbefore, the identification of "locations in the recognized text where silence preceded and/or followed by correctly recognized words" in Heckerman (Col.11, lines 26-30) does not constitute a method of "deriving from the audio data in the scene data feature data representative of audible time-varying acoustic features of the audio data; and populating the database with said feature data" as stated in Claim 27. Although Heckerman does generate text data, there is no scene data.

Comments Regarding Examiner's item 13 and claim 30

The Examiner states (Action page 32) "Wark discloses Graphical user interface software … comprising ..." and cites two passages from Wark and then a third: "The media clip(s) associated with the aforementioned selected icon(s) 804 are played from a selected position and in the desired sequence, in a contiguous fashion as a single media presentation, and continues until the end of the presentation at which point playback stops." (Wark Col. 11, lines 26-30). The Examiner relates this to our third claim portion: "an output program adapted to respond to said control signals by outputting selected synchronized streams of video data and audio data, and to record an input audio stream provided during the said synchronized streams." The fact that "Wark is silent regarding recording of the output of the editor" demonstrates that Wark's teaching is not related to Applicants' invention since in Applicants' invention it is a critical function to record a second input audio signal (namely the user's voice) during the playback of

the original first audio signal synchronized with the video, such second audio signal providing the replacement audio that will be analyzed, edited and the result synchronized with a playback of the video with the edited audio replacing the first audio, which will be suppressed.

Conclusion

Examiner's rejection of Applicants' claims fails to take account of the essential element of "replacing said selected portion" of one audio stream with another edited audio stream. There is no mention in any of the cited references of any means for replacing an audio stream with another or for playing back an audio and video data stream (with moving pictures). Furthermore, there are no means or method of determining timing differences provided in Heckerman. Nor can time differences be generated between an audio stream and a text file which has no timing reference.

More specifically, none of the references cited by the Examiner have the Applicant's invention's means and methods for:

1. Specific, precise and necessary audio, video and other data preparation and storage,

2. Recording and then automated editing of the user(s) voice(s) based on comparing timing differences between detailed acoustic features of **two audio signals: for example, one being** the user's voice and the other being the original actor's or singer's voice (which is in lip-sync with the video images), and

3. the automatic insertion of the user(s) edited (lip-synced) voice(s) at the correct times during playback of the video stream, the playback of other selected audio streams, and the corresponding automatic removal of the original performer's voice during playback to enable the said automatic replacement of the original performer with the users voices.

In view of the foregoing, it is submitted that this application is in condition for allowance. Favorable consideration and prompt allowance of the application are respectfully requested. Applicant believes no fee is due to enter the present Amendment. The Commissioner is hereby authorized to charge any additional filing fees required to Deposit Account No. 061910. The Examiner is invited to telephone the undersigned if the Examiner believes it would be useful to advance prosecution.

Respectfully submitted,

Dated: February 18, 2009

/Charles D. Segelbaum/
Charles D. Segelbaum
Reg. No. 42,138
(612) 492-7115

Fredrikson & Byron, P.A.
200 South Sixth Street, Suite 4000
Minneapolis, MN 55402-1425 USA
Facsimile: (612) 492-7077